Minimum Message Length Inference and Mixture Modelling of Inverse Gaussian Distributions

Daniel F. Schmidt Enes Makalic

Centre for Molecular, Environmental, Genetic & Analytic (MEGA) Epidemiology
School of Population Health
University of Melbourne

25th Australasian Joint Conference on Artificial Intelligence 2012

Content

Problem Description

- We have n items, each with q associated attributes, formed into a matrix

$$\mathbf{Y} = \begin{pmatrix} \mathbf{y}_1 \\ \mathbf{y}_2 \\ \vdots \\ \mathbf{y}_n \end{pmatrix} = \begin{pmatrix} y_{1,1} & y_{1,2} & \cdots & y_{1,q} \\ y_{2,1} & y_{2,2} & \cdots & y_{2,q} \\ \vdots & \vdots & \ddots & \vdots \\ y_{n,1} & y_{n,2} & \cdots & y_{n,q} \end{pmatrix}$$

- Group together, or "cluster", similar items
- A form of unsupervised learning
- Sometimes called intrinsic classification
  $\Rightarrow$ Class labels are learned from the data

Mixture Modelling (1)

- Models data as a mixture of probability distributions

$$p(y_{i,j}; \mathbf{\Phi}) = \sum_{k=1}^{K} \alpha_k p(y_{i,j}; \boldsymbol{\theta}_{k,j})$$

where

  - K is the number of classes
  - $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_K)$ are the mixing (population) weights
  - $\boldsymbol{\theta}_{k,j}$ are the parameters of the distributions
  - $\mathbf{\Phi} = \{K, \boldsymbol{\alpha}, \boldsymbol{\theta}_{1,1}, \ldots, \boldsymbol{\theta}_{K,q}\}$ denotes the complete mixture model

- Has an explicit probabilistic form
  $\Rightarrow$ allows for statistical interpretion

Mixture Modelling (2)

- How is this related to clustering?
- Each class is a cluster
  - Class-specific probability distributions over each attribute
    - e.g., normal, inverse Gaussian, Poisson, etc.
  - Mixing weight is prevalance of the classes in the population
- Measure of similarity of item to class

$$p_k(\mathbf{y}_i) = \prod_{j=1}^{q} p(y_{i,j}; \boldsymbol{\theta}_{k,j})$$

$\Rightarrow$ probability of item's attributes under class distributions

Mixture Modelling (2)

- How is this related to clustering?
- Each class is a cluster
  - Class-specific probability distributions over each attribute
    - e.g., normal, inverse Gaussian, Poisson, etc.
  - Mixing weight is prevalance of the classes in the population
- Measure of similarity of item to class

$$p_k(\mathbf{y}_i) = \prod_{j=1}^{q} p(y_{i,j}; \boldsymbol{\theta}_{k,j})$$

$\Rightarrow$ probability of item's attributes under class distributions

Mixture Modelling (3)

- Membership of items to classes is soft

$$r_{i,k} = \frac{\alpha_k \, p_k(\mathbf{y}_i)}{\sum_{l=1}^{K} \alpha_l \, p_l(\mathbf{y}_i)}$$

- Posterior probability of belonging to class k
  - α_k is *a priori* probability item belongs to class k
  - $p_k(\mathbf{y}_i)$ is probability of data item $\mathbf{y}_i$ under class k

  $\Rightarrow$ Assign to class with highest posterior probability

- Total number of samples in a class is then

$$n_k = \sum_{i=1}^{n} r_{i,k}$$

Mixture Modelling (3)

- Membership of items to classes is soft

$$r_{i,k} = \frac{\alpha_k \, p_k(\mathbf{y}_i)}{\sum_{l=1}^{K} \alpha_l \, p_l(\mathbf{y}_i)}$$

- Posterior probability of belonging to class k
  - α_k is *a priori* probability item belongs to class k
  - $p_k(\mathbf{y}_i)$ is probability of data item $\mathbf{y}_i$ under class k

  $\Rightarrow$ Assign to class with highest posterior probability

- Total number of samples in a class is then

$$n_k = \sum_{i=1}^{n} r_{i,k}$$

MML Mixture Models (1)

- Minimum Message Length goodness-of-fit criterion
  - Popular criterion for mixture modelling
- Based on the idea of compression

- Message length of data is our yardstick, comprised of
  1. Length of codeword needed to state model Φ
     - Number of classes: $I(K)$
     - Relative abundances: $I(\boldsymbol{\alpha})$
     - Parameters for each distribution in each class: $I(\theta_{k,j})$
  2. Length of codeword needed to state data, given model: $I(\mathbf{Y}|\Phi)$

MML Mixture Models (1)

- Minimum Message Length goodness-of-fit criterion
  - Popular criterion for mixture modelling
- Based on the idea of compression

- Message length of data is our yardstick, comprised of
  1. Length of codeword needed to state model $\mathbf{\Phi}$
     - Number of classes: $I(K)$
     - Relative abundances: $I(\boldsymbol{\alpha})$
     - Parameters for each distribution in each class: $I(\boldsymbol{\theta}_{k,j})$
  2. Length of codeword needed to state data, given model: $I(\mathbf{Y}|\mathbf{\Phi})$

MML Mixture Models (2)

- Total message length:

$$I(\mathbf{Y}, \mathbf{\Phi}) = I(K) + I(\boldsymbol{\alpha}) + \sum_{k=1}^{K} \sum_{j=1}^{q} I(\boldsymbol{\theta}_{k,j}) + I(\mathbf{Y}|\mathbf{\Phi})$$

  $\Rightarrow$ balances model complexity against model fit

- Estimate Φ by minimising message length
  - $\hat{\alpha}$ and $\hat{\theta}_{j,k}$ found by expectation-maximisation
  - Find $\hat{K}$ by splitting/merging classes

MML Mixture Models (2)

- Total message length:

$$I(\mathbf{Y}, \mathbf{\Phi}) = I(K) + I(\boldsymbol{\alpha}) + \sum_{k=1}^{K} \sum_{j=1}^{q} I(\boldsymbol{\theta}_{k,j}) + I(\mathbf{Y}|\mathbf{\Phi})$$

  $\Rightarrow$ balances model complexity against model fit

- Estimate $\mathbf{\Phi}$ by minimising message length
  - $\hat{\boldsymbol{\alpha}}$ and $\hat{\boldsymbol{\theta}}_{j,k}$ found by expectation-maximisation
  - Find $\hat{K}$ by splitting/merging classes

Content

Inverse Gaussian Distributions (1)

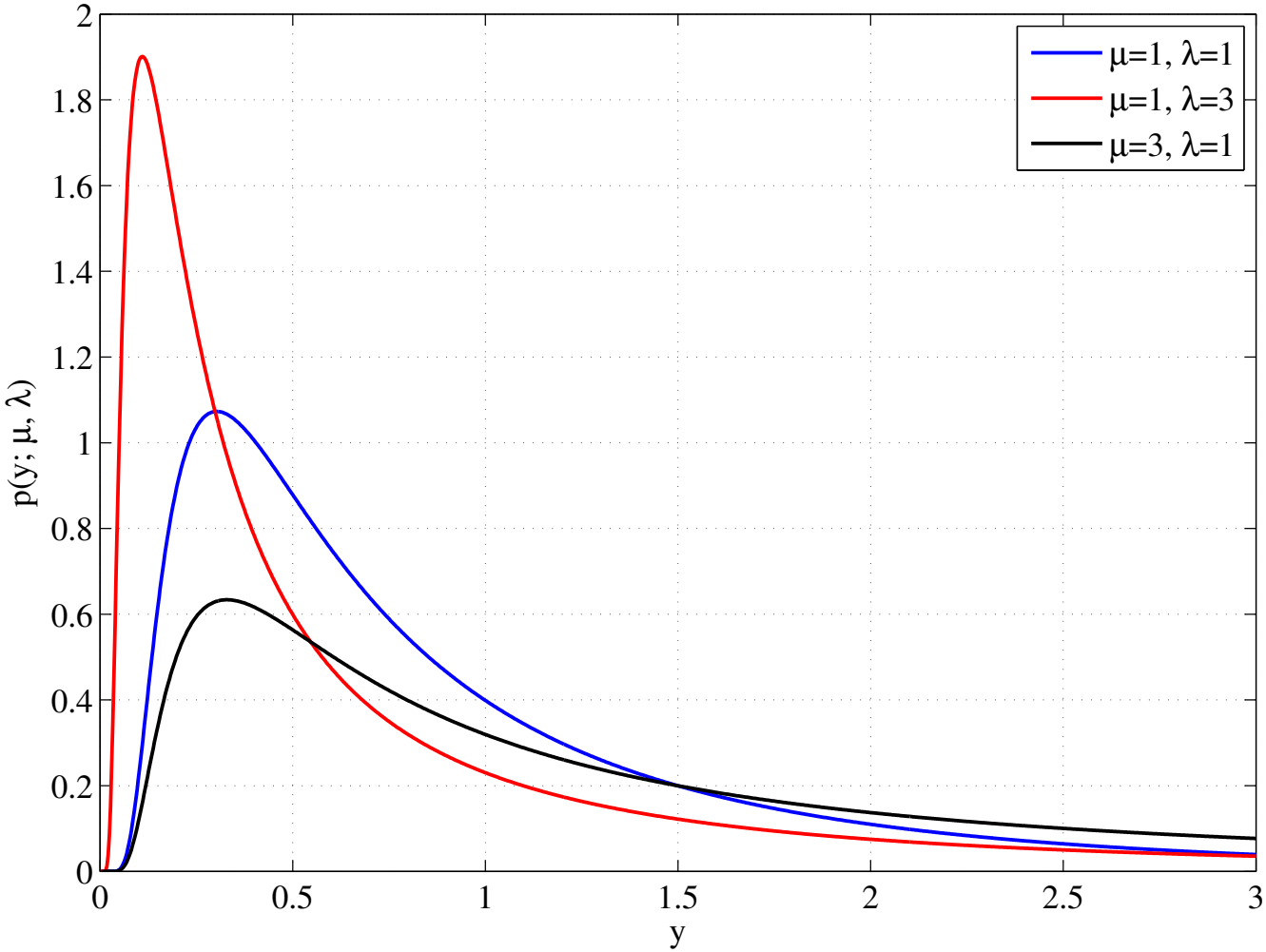
- Distribution for positive, continuous data
- We say $Y_i \sim IG(\mu, \lambda)$ if p.d.f. for $Y_i = y_i$ is

$$p(y_i; \mu, \lambda) = \left(\frac{1}{2\pi\lambda y_i^3}\right)^{\frac{1}{2}} \exp\left(-\frac{(y_i - \mu)^2}{2\mu^2\lambda y_i}\right),$$

  where
  - $\mu > 0$ is the mean parameter
  - $\lambda > 0$ is the inverse-shape parameter
- Suitable for positively skewed data
- Derive the message length formula for use in mixture modelling

Inverse Gaussian Distributions (2)

- Example of inverse Gaussian distributions



MML Inference of Inverse Gaussians (1)

- Use Wallace–Freeman approximation
- Bayesian; we chose uninformative priors

$$\pi(\mu, \lambda) \propto \frac{1}{\lambda \mu^{\frac{3}{2}}}$$

- Message length component for use in mixture models

$$I(\boldsymbol{\theta}_{k,j}) = \log n_k - \frac{1}{2} \log \hat{\lambda}_{k,j} + \log \left(\frac{2\sqrt{2} a_j}{\sqrt{b_j}} \right)$$

where

- $\hat{\lambda}_{k,j}$ is the MML estimate of λ for class k and variable j
- n_k is number of samples in class k
- a_j, b_j are hyper-parameters

- Details may be found in the paper

MML Inference of Inverse Gaussians (1)

- Use Wallace–Freeman approximation
- Bayesian; we chose uninformative priors

$$\pi(\mu, \lambda) \propto \frac{1}{\lambda \mu^{\frac{3}{2}}}$$

- Message length component for use in mixture models

$$I(\boldsymbol{\theta}_{k,j}) = \log n_k - \frac{1}{2} \log \hat{\lambda}_{k,j} + \log \left(\frac{2\sqrt{2} a_j}{\sqrt{b_j}} \right)$$

where

- $\hat{\lambda}_{k,j}$ is the MML estimate of λ for class k and variable j
- n_k is number of samples in class k
- a_j, b_j are hyper-parameters

- Details may be found in the paper

MML Inference of Inverse Gaussians (2)

- Let $\mathbf{y} = (y_1, \ldots, y_n)$ be data from an inverse Gaussian
- Define sufficient statistics

$$S_1 = \sum_{i=1}^{n} y_i, \quad S_2 = \sum_{i=1}^{n} \frac{1}{y_i},$$

- Compare maximum likelihood estimates

$$\hat{\mu}_{\mathrm{ML}} = \frac{S_1}{n}, \quad \hat{\lambda}_{\mathrm{ML}} = \frac{S_1 S_2 - n^2}{n S_1}$$

to minimum message length estimates

$$\hat{\mu}_{87} = \frac{S_1}{n}, \quad \hat{\lambda}_{87} = \frac{S_1 S_2 - n^2}{(n-1) S_1}$$

- MML estimates:
  1. Are Unbiased
  2. Strictly dominate ML estimates in terms of KL risk

MML Inference of Inverse Gaussians (2)

- Let $\mathbf{y} = (y_1, \ldots, y_n)$ be data from an inverse Gaussian
- Define sufficient statistics

$$S_1 = \sum_{i=1}^{n} y_i, \quad S_2 = \sum_{i=1}^{n} \frac{1}{y_i},$$

- Compare maximum likelihood estimates

$$\hat{\mu}_{\mathrm{ML}} = \frac{S_1}{n}, \quad \hat{\lambda}_{\mathrm{ML}} = \frac{S_1 S_2 - n^2}{n S_1}$$

to minimum message length estimates

$$\hat{\mu}_{87} = \frac{S_1}{n}, \quad \hat{\lambda}_{87} = \frac{S_1 S_2 - n^2}{(n-1) S_1}$$

- MML estimates:
  1. Are Unbiased
  2. Strictly dominate ML estimates in terms of KL risk

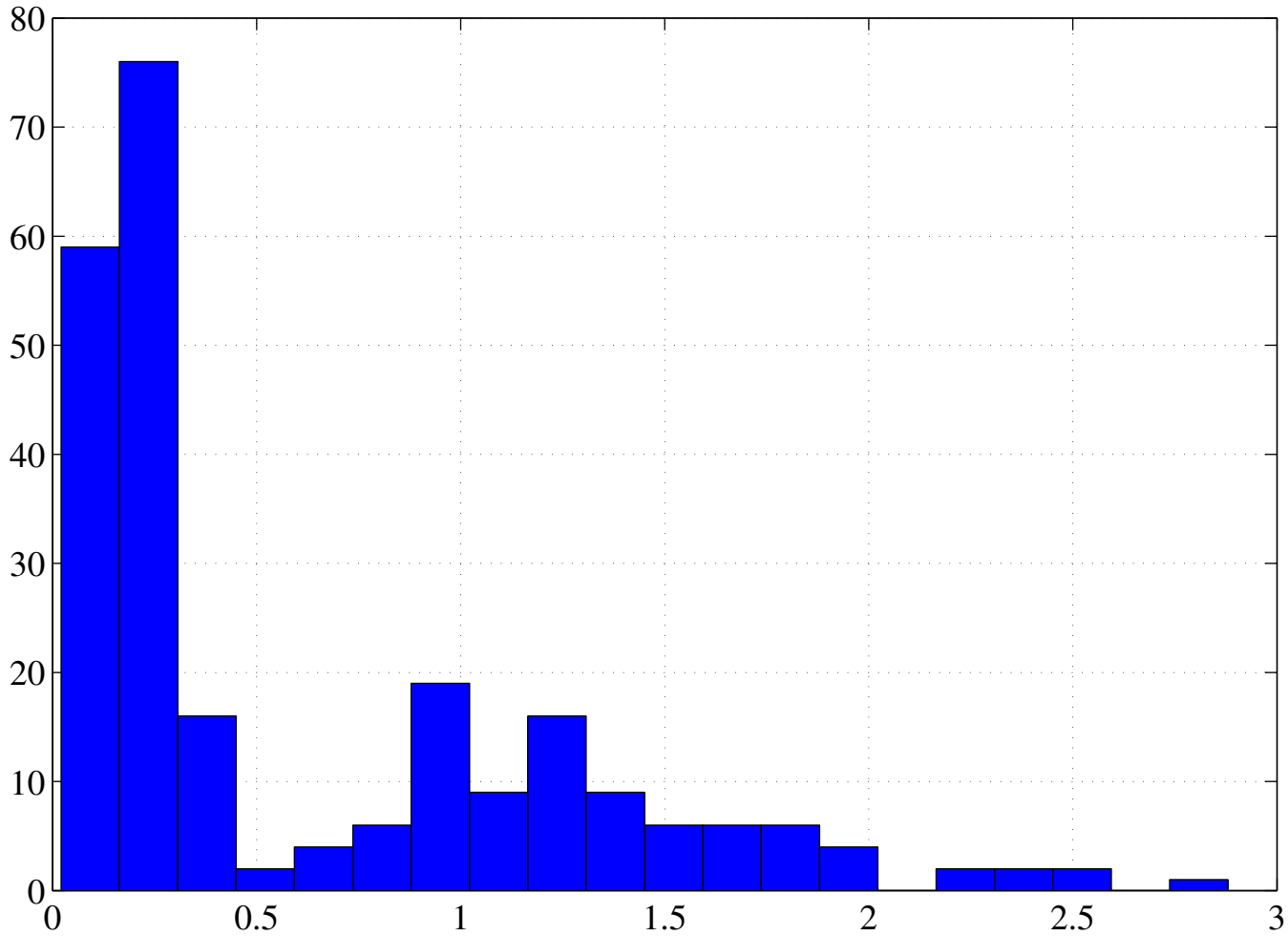
Content

Example (1)

- Compared inverse Gaussian mixture models against standard Gaussian mixture models
- Used several well known, real, datasets
  1. "Enzyme"
  2. "Acidity"
  3. "Galaxy"
- Results shown for "enzyme"
  - $n = 245$ samples
- See paper for "acidity" and "galaxy" results

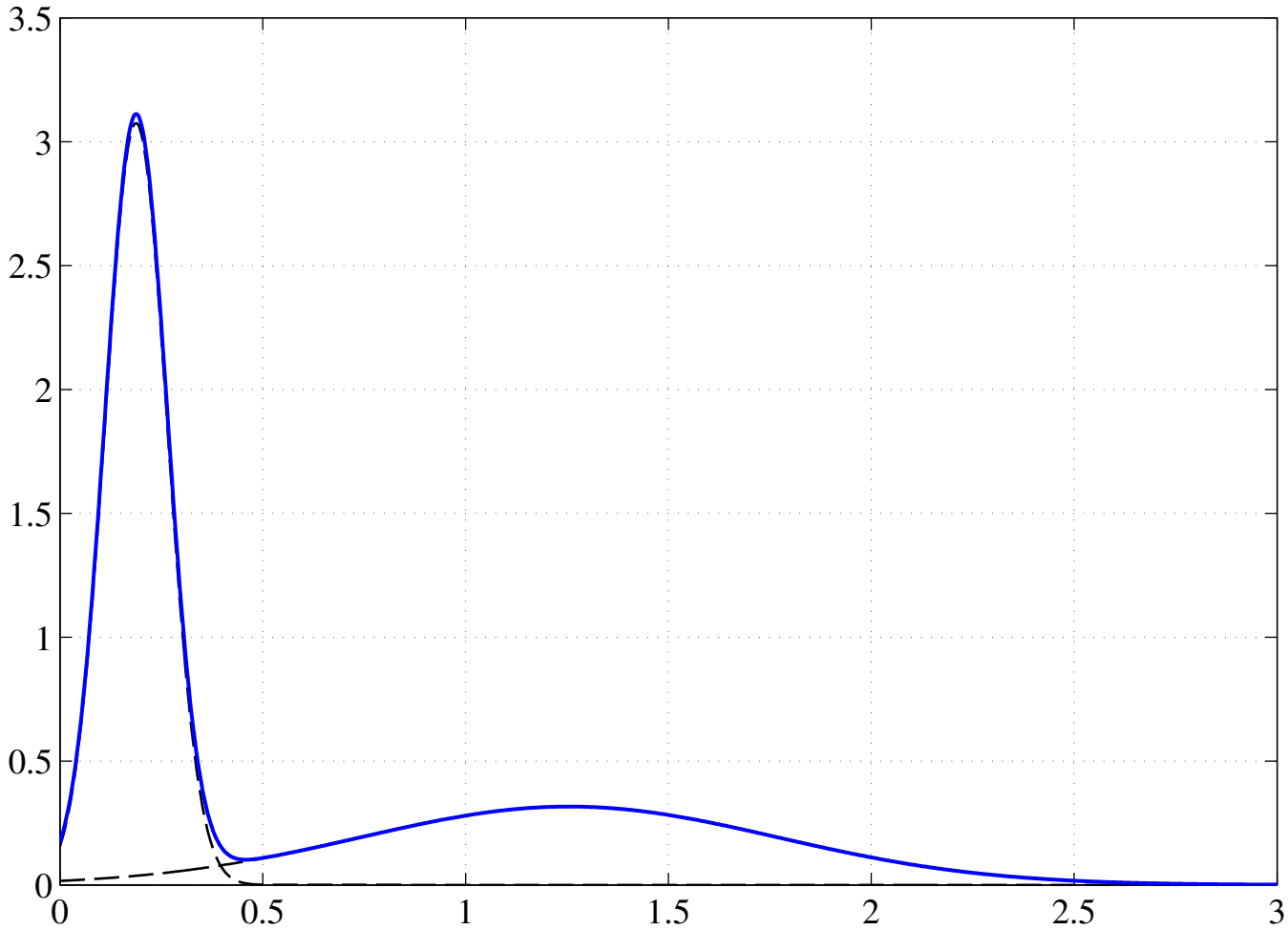
Example (2)

- Histogram of "enzyme" data



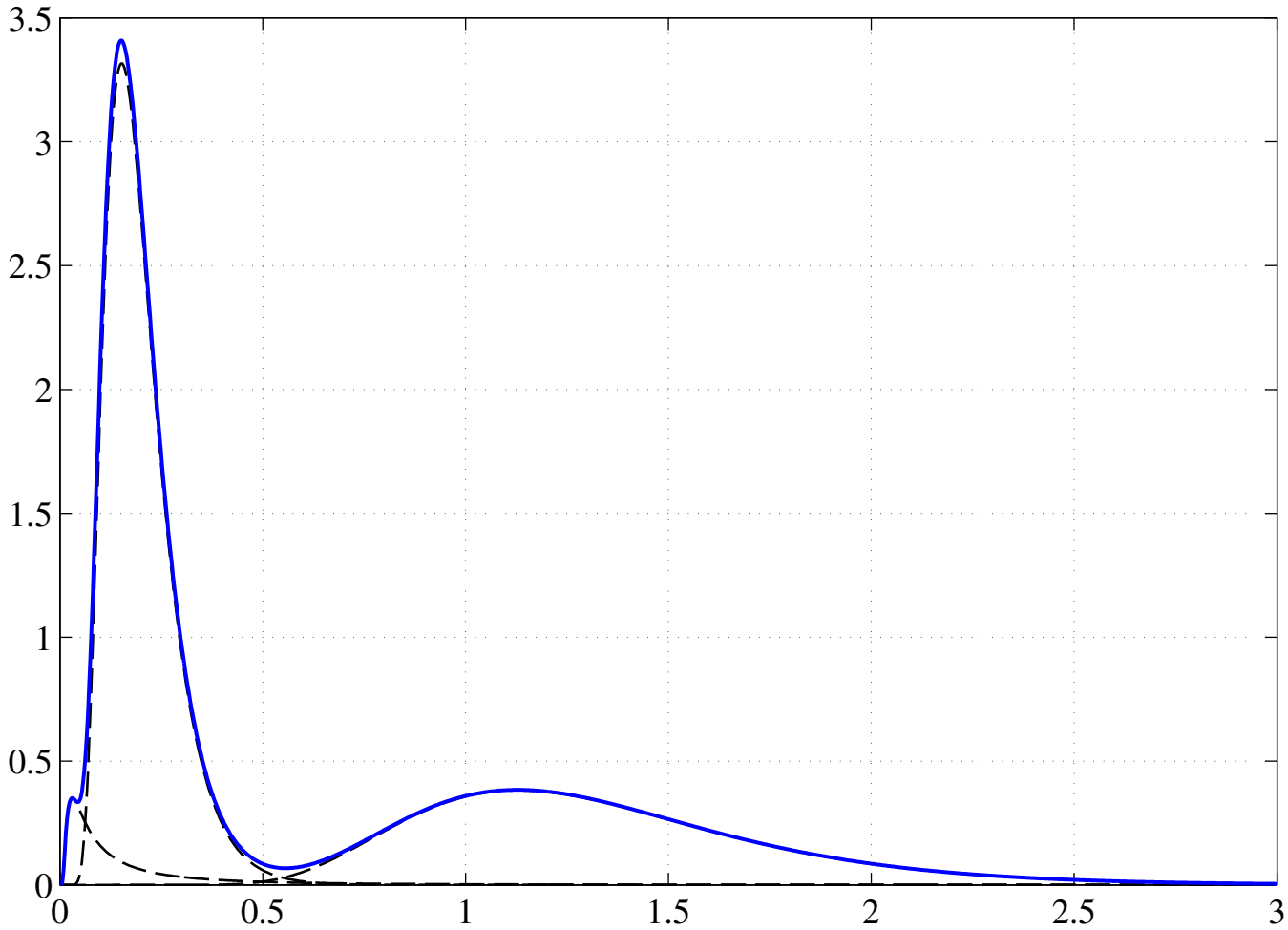
Example (3)

- Gaussian mixture model ($K = 2$, $I = 86.19$)



Example (4)

- Inverse Gaussian mixture model ($K = 3$, $I = 69.34$)



References

- Wallace, C. S., Boulton, D. M. "An information measure for classification". *Computer Journal*, 1968, Vol. 11, pp. 185-194
- Wallace, C. S., Dowe, D. L. "MML mixture modelling of multi-state, Poisson, von Mises circular and Gaussian distributions". *Proceedings of the 6th International Workshop on Artificial Intelligence and Statistics*, 1997, pp. 529-536
- Wallace, C. S. "Intrinsic Classification of Spatially Correlated Data". *The Computer Journal*, 1998, Vol. 41, pp. 602-611
- Wallace, C. S., Dowe, D. L., "MML clustering of multi-state, Poisson, von Mises circular and Gaussian distributions". *Statistics and Computing*, 2000, Vol. 10, pp. 73-83
- Wallace, C. S. "Statistical and Inductive Inference by Minimum Message Length", *Springer*, 2005